CLAIMS

1. A method of identifying one or more portions of a document, the method comprising:

identifying a plurality of visual blocks in the document;

detecting one or more separators between the visual blocks of the plurality of visual blocks; and

constructing, based at least in part on the plurality of visual blocks and the one or more separators, a content structure for the document, wherein the content structure identifies the different visual blocks as different portions of semantic content of the document.

2. A method as recited in claim 1, wherein the document comprises a web page.

3. A method as recited in claim 1, wherein the document is described by a tree structure having a plurality of nodes, and wherein identifying the plurality of visual blocks in the document comprises:

identifying a group of candidate nodes of the plurality of nodes;

for each node in the group of candidate nodes:

determining whether the node can be divided, and

if the node cannot be divided, then identifying the node as representing a visual block.

4. A method as recited in claim 3, wherein if the node cannot be divided, then setting a degree of coherence for the visual block represented by the node.

5. A method as recited in claim 3, wherein if the node cannot be divided, then removing the node from the group of candidate nodes.

6. A method as recited in claim 3, wherein determining whether the node can be divided comprises determining that the node can be divided if the node has a child node with <HR> HyperText Markup Language (HTML) tag.

7. A method as recited in claim 3, wherein determining whether the node can be divided comprises determining that the node can be divided if a background color of the node is different from a background color of a child of the node.

8. A method as recited in claim 3, further comprising checking whether the node has a child having a width and height greater than zero, and if the node has no child having a width and height greater than zero then removing the node from the group of candidate nodes.

9. A method as recited in claim 3, wherein determining whether the node can be divided comprises determining that the node can be divided if a size of the node is at least a threshold amount greater than a sum of sizes of children nodes of the node.

10. A method as recited in claim 3, wherein determining whether the node can be divided comprises determining that the node can be divided if the node has multiple successive children nodes each having a
 HyperText Markup Language (HTML) tag.

11. A method as recited in claim 1, wherein the document is described by a tree structure having a plurality of nodes, and wherein identifying the plurality of visual blocks in the document comprises identifying different visual blocks based at least in part on HyperText Markup Language (HTML) tags of the plurality of nodes.

12. A method as recited in claim 1, wherein the document is described by a tree structure having a plurality of nodes, and wherein identifying the plurality of visual blocks in the document comprises identifying different visual blocks based at least in part on background colors of the plurality of nodes.

13. A method as recited in claim 1, wherein the document is described by a tree structure having a plurality of nodes, and wherein identifying the plurality of visual blocks in the document comprises identifying different visual blocks based at least in part on whether the plurality of nodes include text and the sizes of the plurality of nodes.

14. A method as recited in claim 1, wherein detecting the one or more separators comprises:

detecting one or more horizontal separators between the visual blocks; and

detecting one or more vertical separators between the visual blocks.

15. A method as recited in claim 1, wherein detecting the one or more separators comprises:

initializing a separator list that includes one or more possible separators between the visual blocks;

analyzing, for each of the visual blocks, whether the visual block overlaps a separator of the separator list, and if so how the visual block overlaps the separator; and

determining how to treat the separator based on whether the visual block overlaps the separator, and if so how the visual block overlaps the separator.

16. A method as recited in claim 15, further comprising determining to split the separator into multiple separators if the visual block is contained in the separator.

17. A method as recited in claim 15, further comprising determining, if the visual block crosses the separator, to modify parameters of the separator so that the visual block no longer crosses the separator.

18. A method as recited in claim 17, wherein the modification comprises reducing the height of the separator if the separator is a horizontal separator.

19. A method as recited in claim 17, wherein the modification comprises reducing the width of the separator if the separator is a vertical separator.

20. A method as recited in claim 15, further comprising determining to remove the separator from the separator list if the visual block covers the separator.

21. A method as recited in claim 1, further comprising assigning, to each of the one or more separators, a weight based on characteristics of visual blocks on either side of the separator.

22. A method as recited in claim 21, wherein assigning the weight comprises assigning the weight based on a distance between two visual blocks on either side of the separator.

23. A method as recited in claim 21, wherein assigning the weight comprises assigning the weight based on whether the separator is at a same position as an <HR> HTML tag.

24. A method as recited in claim 21, wherein assigning the weight comprises assigning the weight based on a font size used in two visual blocks on either side of the separator.

25. A method as recited in claim 21, wherein assigning the weight comprises assigning the weight based on a background color used in two visual blocks on either side of the separator.

26. A method as recited in claim 1, further comprising:

checking whether each of the plurality of visual blocks satisfies a degree of coherence threshold; and

for each of the plurality of visual blocks that does not satisfy the degree of coherence threshold, identifying a new plurality of visual blocks in the visual block, and repeating the detecting and constructing using the new plurality of visual blocks.

27. A method as recited in claim 1, wherein constructing the content structure comprises:

generating one or more virtual blocks based on the plurality of visual blocks; and

including, in the content structure, the one or more virtual blocks.

28. A method as recited in claim 27, wherein generating the one or more virtual blocks comprises generating the one or more virtual blocks by combining two visual blocks of the plurality of visual blocks.

29. A method as recited in claim 27, further comprising:

determining a degree of coherence value for each of the one or more virtual blocks.

30. A method as recited in claim 29, wherein determining the degree of coherence value for a virtual block comprises determining the degree of coherence value for the virtual block based at least in part on a weight of a separator between two visual blocks used to generate the virtual block.

31. One or more computer readable media having stored thereon a plurality of instructions that, when executed by one or more processors of a device, causes the one or more processors to:

identify visual blocks in a document;

detect visual separators between the visual blocks; and

construct, based at least in part on the visual blocks and the visual separators, a content structure for the document that identifies regions of the document that represent semantic content of the document.

32. One or more computer readable media as recited in claim 31, wherein the document is described by a tree structure having a plurality of nodes, and wherein the instructions that cause the one or more processors to identify visual blocks in the document comprise instructions that cause the one or more processors to:

identify a group of candidate nodes of the plurality of nodes;

for each node in the group of candidate nodes:

determine whether the node can be divided, and

if the node cannot be divided, then identify the node as representing a visual block.

33. One or more computer readable media as recited in claim 31, wherein the instructions that cause the one or more processors to detect visual separators comprise instructions that cause the one or more processors to:

detect one or more horizontal separators between the visual blocks; and

detect one or more vertical separators between the visual blocks.

34. One or more computer readable media as recited in claim 31, wherein the instructions that cause the one or more processors to detect visual separators comprise instructions that cause the one or more processors to:

initialize a separator list that includes one or more possible visual separators between the visual blocks;

analyze, for each of the visual blocks, whether the visual block overlaps a separator of the separator list, and if so how the visual block overlaps the separator; and

determine how to treat the separator based on whether the visual block overlaps the separator, and if so how the visual block overlaps the separator.

35. One or more computer readable media as recited in claim 31, wherein the instructions further cause the one or more processors to:

check whether each of the visual blocks satisfies a degree of coherence threshold; and

for each of the visual blocks that does not satisfy the degree of coherence threshold, identify new visual blocks in the visual block, and repeat the detection and construction using the new visual blocks.

36. A method of searching a plurality of documents, the method comprising:

receiving query criteria corresponding to a query;

accessing a plurality of blocks corresponding to the plurality of documents, wherein different blocks of the plurality of blocks correspond to different documents of the plurality of documents, wherein the plurality of blocks have been obtained by visually segmenting each of the plurality of documents;

generating rankings for one or more of the plurality of blocks based at least in part on how well the blocks match the query criteria;

generating rankings for one or more of the plurality of documents, wherein the ranking of each of the plurality of documents is based at least in part on the rankings of the multiple blocks corresponding to the document; and

returning an indication of at least one of the one or more ranked documents.

37. A method as recited in claim 36, wherein each of the plurality of documents comprises a web page.

38. A method as recited in claim 36, wherein generating the ranking for one of the plurality of documents comprises:

identifying the rankings of each of the multiple blocks corresponding to the one document;

selecting, as the ranking for the one document, the highest ranking of the identified rankings.

39. A method as recited in claim 36, wherein generating the ranking for one of the plurality of documents comprises:

identifying the rankings of each of the multiple blocks corresponding to the one document; and

combining the identified rankings to form a ranking for the one document.

40. A method as recited in claim 39, wherein the combining comprises averaging the identified rankings.

41. A method as recited in claim 36, wherein the visually segmenting a document comprises:

identifying a plurality of visual blocks in the document;

detecting one or more separators between the visual blocks of the plurality of visual blocks; and

constructing, based at least in part on the plurality of visual blocks and the one or more separators, a content structure for the document, wherein the content structure identifies the different visual blocks as different portions of semantic content of the document, and wherein the different visual blocks are the blocks of the plurality of blocks that correspond to the document.

42. A method as recited in claim 41, wherein the document is described by a tree structure having a plurality of nodes, and wherein identifying the plurality of visual blocks in the document comprises:

identifying a group of candidate nodes of the plurality of nodes;

for each node in the group of candidate nodes:

determining whether the node can be divided, and

if the node cannot be divided, then identifying the node as representing a visual block.

43. One or more computer readable media having stored thereon a plurality of instructions that, when executed by one or more processors of a device, causes the one or more processors to:

receive a query including one or more search terms;

rank a plurality of blocks based on how well the plurality of blocks matches the one or more search terms, wherein each of the plurality of blocks is part of one document of a plurality of documents, and wherein each of the plurality of blocks is obtained by visual segmentation of one of the plurality of documents;

for each of the plurality of documents, rank the document based at least in part on the rankings of the blocks that are part of the document; and

return, in response to the query, an indication of the rankings of one or more of the plurality of documents.

44. One or more computer readable media as recited in claim 43, wherein the instructions that cause the one or more processors to rank the document comprise instructions that cause the one or more processors to:

identify the ranking for each block that is part of the document;

select, as the ranking for the document, the highest ranking of the identified rankings.

45. One or more computer readable media as recited in claim 43, wherein the instructions that cause the one or more processors to rank the document comprise instructions that cause the one or more processors to:

identify the ranking for each block that is part of the document;

combine the rankings for each block to generate a ranking for the document.

46. One or more computer readable media as recited in claim 43, wherein the visual segmentation of a document comprises:

identifying a plurality of visual blocks in the document;

detecting one or more separators between the visual blocks of the plurality of visual blocks; and

constructing, based at least in part on the plurality of visual blocks and the one or more separators, a content structure for the document, wherein the content structure identifies the different visual blocks as different portions of semantic

content of the document, and wherein the different visual blocks are the blocks of the plurality of blocks that are part of the document.

47. One or more computer readable media as recited in claim 46, wherein the document is described by a tree structure having a plurality of nodes, and wherein identifying the plurality of visual blocks in the document comprises:

identifying a group of candidate nodes of the plurality of nodes;

for each node in the group of candidate nodes:

determining whether the node can be divided, and

if the node cannot be divided, then identifying the node as representing a visual block.

48. A method of searching a plurality of web pages, the method comprising:

receiving a request to search the plurality of web pages;

generating a first set of rankings for a subset of the plurality of web pages based on the request;

generating a second set of rankings for the subset of web pages by visually segmenting each web page in the subset of web pages; and

obtaining, based at least in part on the second set of rankings, a final set of rankings for the subset of web pages.

49. A method as recited in claim 48, wherein obtaining the final set of rankings comprises using, as the final set of rankings, the second set of rankings.

50. A method as recited in claim 48, wherein obtaining the final set of rankings comprises selecting, as the final ranking for a web page, the higher ranking of the ranking of the web page in the first set and the ranking of the web page in the second set.

51. A method as recited in claim 48, wherein obtaining the final set of rankings comprises averaging, to obtain the final ranking for a web page, the ranking of the web page in the first set and the ranking of the web page in the second set.

52. A method as recited in claim 48, wherein visually segmenting a web page comprises:

identifying a plurality of visual blocks in the web page;

detecting one or more separators between the visual blocks of the plurality of visual blocks; and

constructing, based at least in part on the plurality of visual blocks and the one or more separators, a content structure for the web page, wherein the content structure identifies the different visual blocks as different portions of semantic content of the web page.

53. A method as recited in claim 52, wherein the web page is described by a tree structure having a plurality of nodes, and wherein identifying the plurality of visual blocks in the web page comprises:

identifying a group of candidate nodes of the plurality of nodes;

for each node in the group of candidate nodes:

determining whether the node can be divided, and

if the node cannot be divided, then identifying the node as representing a visual block.

54. One or more computer readable media having stored thereon a plurality of instructions that, when executed by one or more processors of a device, causes the one or more processors to:

generate first rankings for a plurality of documents based on how well the plurality of documents match search criteria;

generate second rankings for the plurality of documents by visually segmenting each of the plurality of documents; and

generate final rankings for the plurality of documents based at least in part on the second rankings.

55. One or more computer readable media as recited in claim 54, wherein the instructions that cause the one or more processors to generate final rankings comprise instructions that cause the one or more processors to use, as the final rankings, the second rankings.

56. One or more computer readable media as recited in claim 54, wherein the instructions that cause the one or more processors to generate final rankings comprise instructions that cause the one or more processors to select, as a final ranking for a document of the plurality of documents, whichever ranking of the first ranking for the document and the second ranking of the document is higher.

57. One or more computer readable media as recited in claim 54, wherein the instructions that cause the one or more processors to generate final rankings comprise instructions that cause the one or more processors to generate a final ranking for a document of the plurality of documents by averaging the first ranking of the document and the second ranking of the document.

58. One or more computer readable media as recited in claim 54, wherein the instructions that cause the one or more processors to visually segment a document comprise instructions that cause the one or more processors to:

identify a plurality of visual blocks in the document;

detect one or more separators between the visual blocks of the plurality of visual blocks; and

construct, based at least in part on the plurality of visual blocks and the one or more separators, a content structure for the document, wherein the content structure identifies the different visual blocks as different portions of semantic content of the document.

59. One or more computer readable media as recited in claim 58, wherein the document is described by a tree structure having a plurality of nodes, and wherein the instructions that cause the one or more processors to identify the plurality of visual blocks in the document comprise instructions that cause the one or more processors to:

identify a group of candidate nodes of the plurality of nodes;

for each node in the group of candidate nodes:

determine whether the node can be divided, and

if the node cannot be divided, then identify the node as representing a visual block.

60. A method of searching a plurality of documents, the method comprising:

receiving a request to search the plurality of documents, wherein the request includes query criteria;

identifying a subset of the plurality of documents based on the query criteria;

identifying, for each of the subset of documents, a plurality of blocks by visually segmenting the document;

expanding, based on the content of the plurality of blocks, the query criteria; and

identifying a second subset of the plurality of documents based on the expanded query criteria.

61. A method as recited in claim 60, returning, in response to the request, identifiers of the second subset of documents.

62. A method as recited in claim 60, ranking each document of the second subset of the plurality of documents; and

returning, in response to the request, identifiers of the second subset of documents and an indication of the ranking of each document of the second subset of documents.

63. A method as recited in claim 60, wherein the visually segmenting the document comprises:

identifying a plurality of visual blocks in the document;

detecting one or more separators between the visual blocks of the plurality of visual blocks; and

constructing, based at least in part on the plurality of visual blocks and the one or more separators, a content structure for the document, wherein the content structure identifies the different visual blocks as different portions of semantic content of the document, and wherein the different visual blocks are the plurality of blocks for the document.

64. A method as recited in claim 63, wherein the document is described by a tree structure having a plurality of nodes, and wherein identifying the plurality of visual blocks in the document comprises:

identifying a group of candidate nodes of the plurality of nodes;

for each node in the group of candidate nodes:

determining whether the node can be divided, and

if the node cannot be divided, then identifying the node as representing a visual block.

65. One or more computer readable media having stored thereon a plurality of instructions that, when executed by one or more processors of a device, causes the one or more processors to:

receive one or more search terms;

identify a plurality of documents that satisfy the one or more search terms;

perform vision-based document segmentation on each of the plurality of documents to identify blocks of each of the plurality of documents;

generate a rank for each of the identified blocks based on how well the block matches the one or more search terms;

derive one or more expansion terms from one or more of the identified blocks; and

identify another plurality of documents that satisfy the one or more search terms and the expansion terms.

66. One or more computer readable media as recited in claim 65, wherein the instructions that cause the one or more processors to derive the one or more expansion terms cause the one or more processors to derive the one or more expansion terms from a group of top-ranked identified blocks.

67. One or more computer readable media as recited in claim 65, wherein the instructions that cause the one or more processors to perform vision-based document segmentation comprise instructions that cause the one or more processors to:

identify a plurality of visual blocks in the document;

detect one or more separators between the visual blocks of the plurality of visual blocks; and

construct, based at least in part on the plurality of visual blocks and the one or more separators, a content structure for the document, wherein the content structure identifies the different visual blocks as different portions of semantic content of the document, and wherein the different visual blocks are the blocks of the document.

68. A system comprising:

a visual block extractor to extract visual blocks from a document;

a visual separator detector coupled to receive the extracted visual blocks and detect, based on the extracted visual blocks, one or more visual separators between the extracted visual blocks; and

a content structure constructor coupled to receive the extracted visual blocks and the detected visual separators, and to use the extracted visual blocks and the detected visual separators to construct a content structure for the document.

69. A system as recited in claim 68, further comprising:

a document retrieval module to retrieve documents from a plurality of documents based at least in part on the content structure constructed for one or more of the plurality of documents.

70. A system as recited in claim 68, wherein the document is described by a tree structure having a plurality of nodes, and wherein the visual block extractor is to extract visual blocks from the document by:

identifying a group of candidate nodes of the plurality of nodes;

for each node in the group of candidate nodes:

determining whether the node can be divided, and

if the node cannot be divided, then identifying the node as representing a visual block.

71. A system as recited in claim 68, wherein the visual separator detector is to detect one or more horizontal separators between the visual blocks and one or more vertical separators between the visual blocks.

72. A system as recited in claim 68, wherein the visual separator detector is to detect the one or more separators by:

initializing a separator list that includes one or more possible separators between the visual blocks;

analyzing, for each of the visual blocks, whether the visual block overlaps a separator of the separator list, and if so how the visual block overlaps the separator; and

determining how to treat the separator based on whether the visual block overlaps the separator, and if so how the visual block overlaps the separator.

73. A system as recited in claim 68, wherein the content structure constructor is further to:

check whether each of the plurality of visual blocks satisfies a degree of coherence threshold; and

for each of the plurality of visual blocks that does not satisfy the degree of coherence threshold, return the visual block to the visual block extractor to have a new plurality of visual blocks extracted from the visual block, and further to have the visual separator detector detect one or more visual separators using the new plurality of visual blocks.

74. A system comprising:

means for identifying a plurality of visual blocks in the document;

means for detecting one or more separators between the visual blocks of the plurality of visual blocks; and

means for constructing, based at least in part on the plurality of visual blocks and the one or more separators, a content structure for the document, wherein the content structure identifies the different visual blocks as different portions of semantic content of the document.

75. A system as recited in claim 74, wherein the document is described by a tree structure having a plurality of nodes, and wherein the means for identifying the plurality of visual blocks in the document comprises:

means for identifying a group of candidate nodes of the plurality of nodes;

for each node in the group of candidate nodes:

means for determining whether the node can be divided, and

means for identifying, if the node cannot be divided, the node as representing a visual block.